Tools For Computational Biology – Lecture 17

Introduction to (bulk) RNA-seq

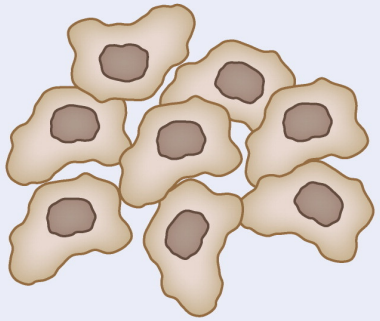
Overview

- A very brief introduction to RNA-seq

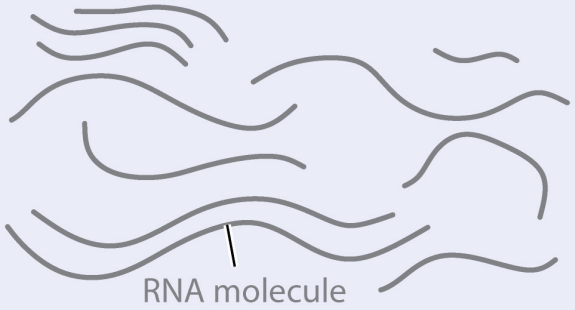
- Bulk RNA-seq analysis – R markdown
  - Quality controls and checks
  - Count matrices
  - Differential analysis
  - Gene Set Enrichment Analysis / Gene Ontology Analysis
  - Visualization

RNA-seq

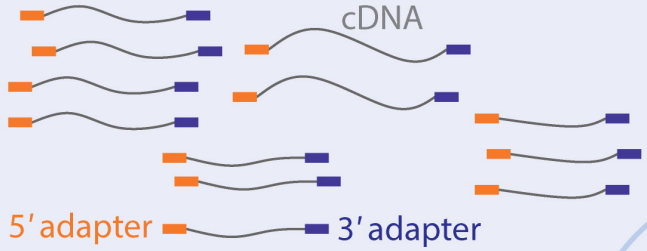
Sample of interest



Total RNA extraction and target enrichment



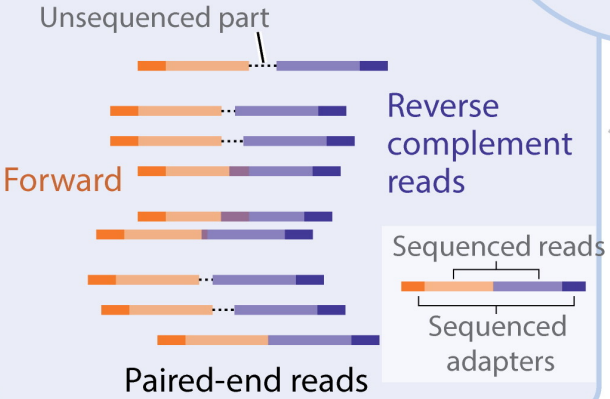
Fragment, reverse transcribe, ligate adapters, amplify



cDNA library

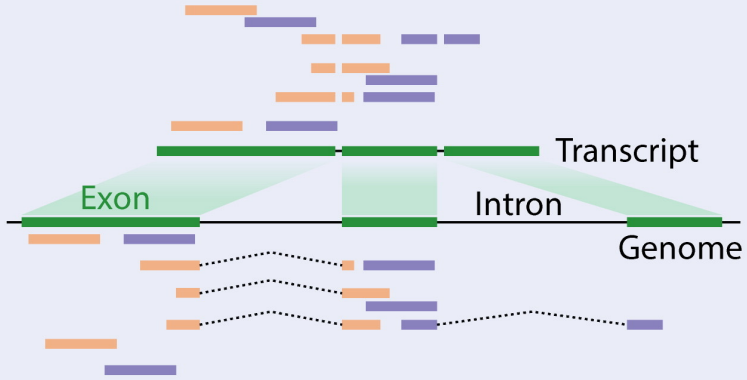
Sequencing machine

Sequencing



Paired-end reads

Transcriptome/genome mapping

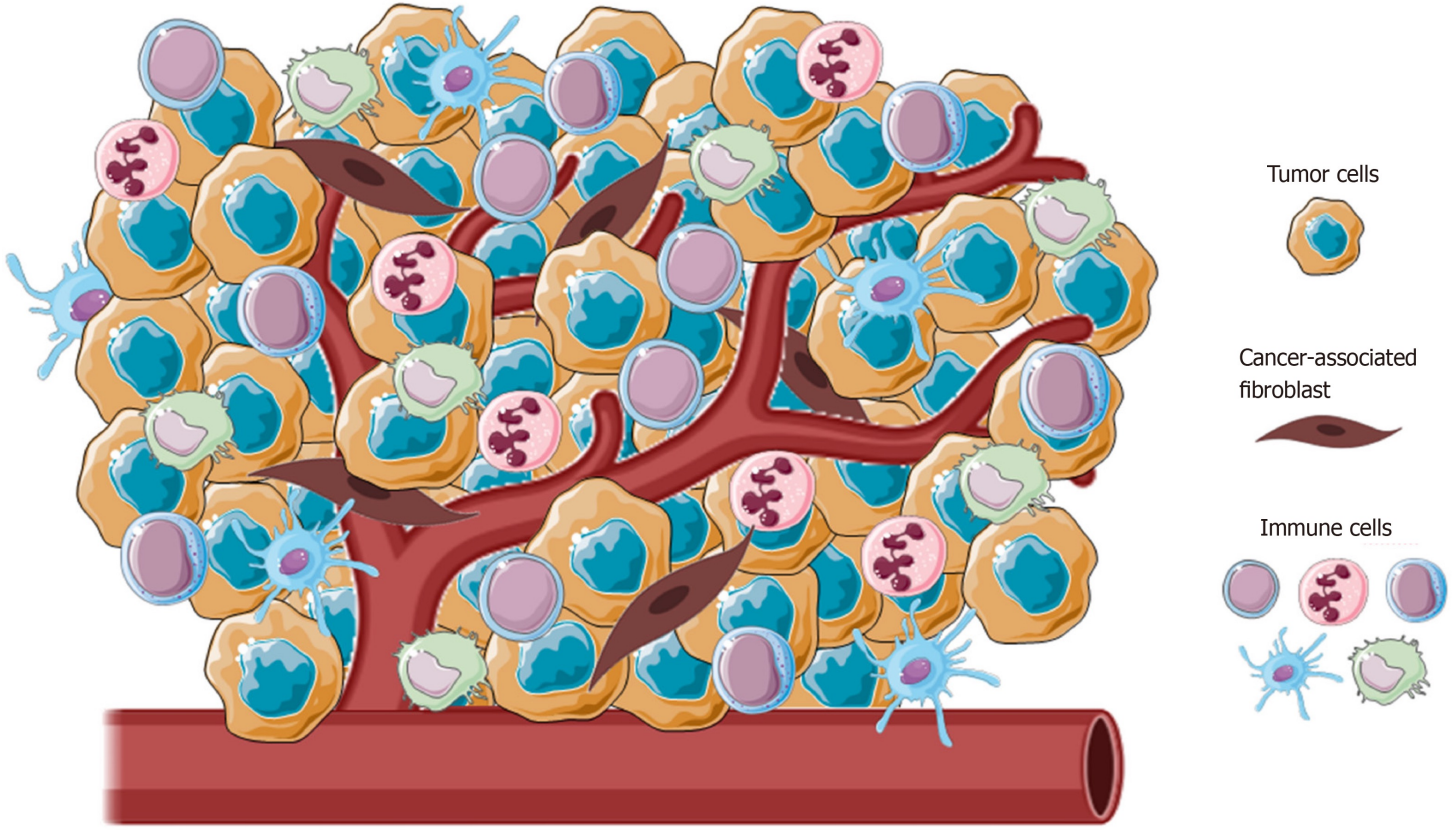


Data analysis

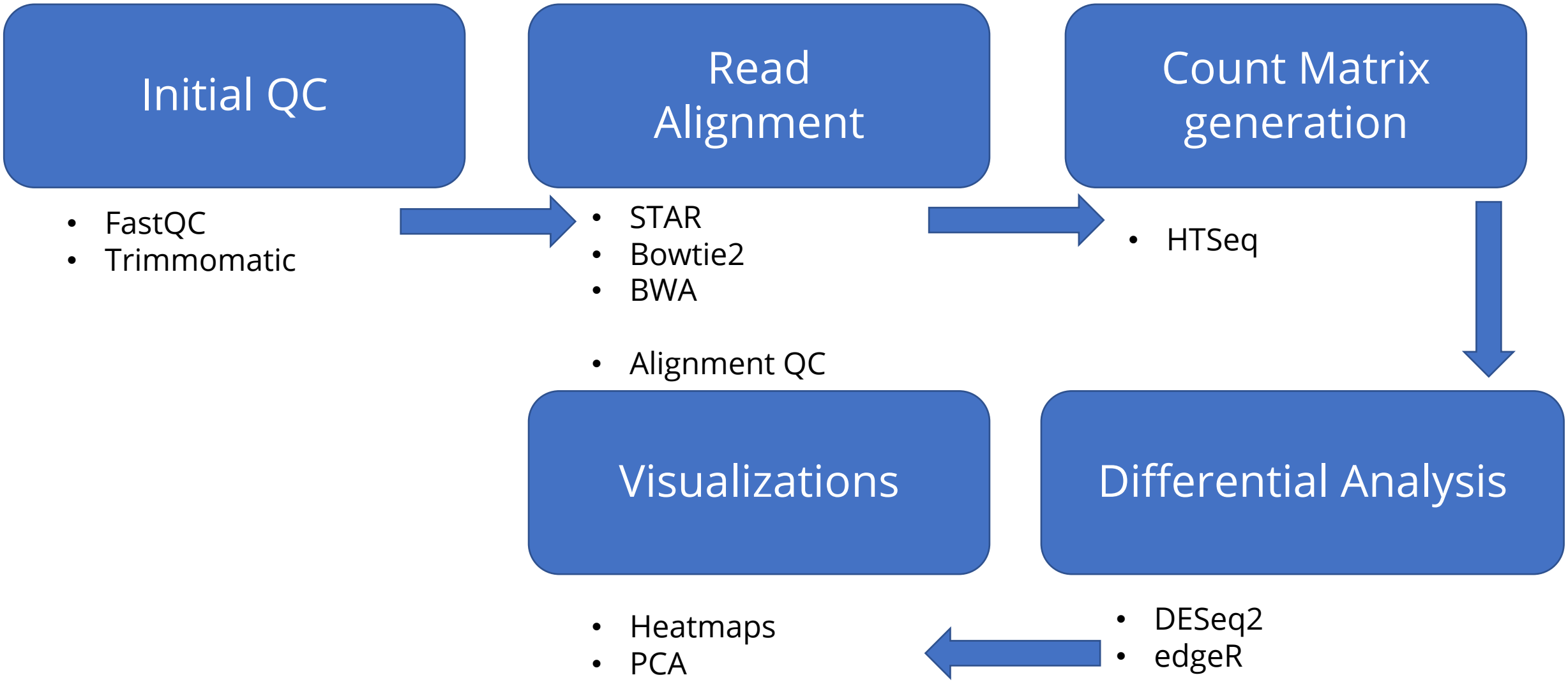
- Differential expression
- Variant calling
- Annotation
- Novel transcript discovery
- RNA editing

Van den Berge K, et al. 2019.
Annu. Rev. Biomed. Data Sci. 2:139–73

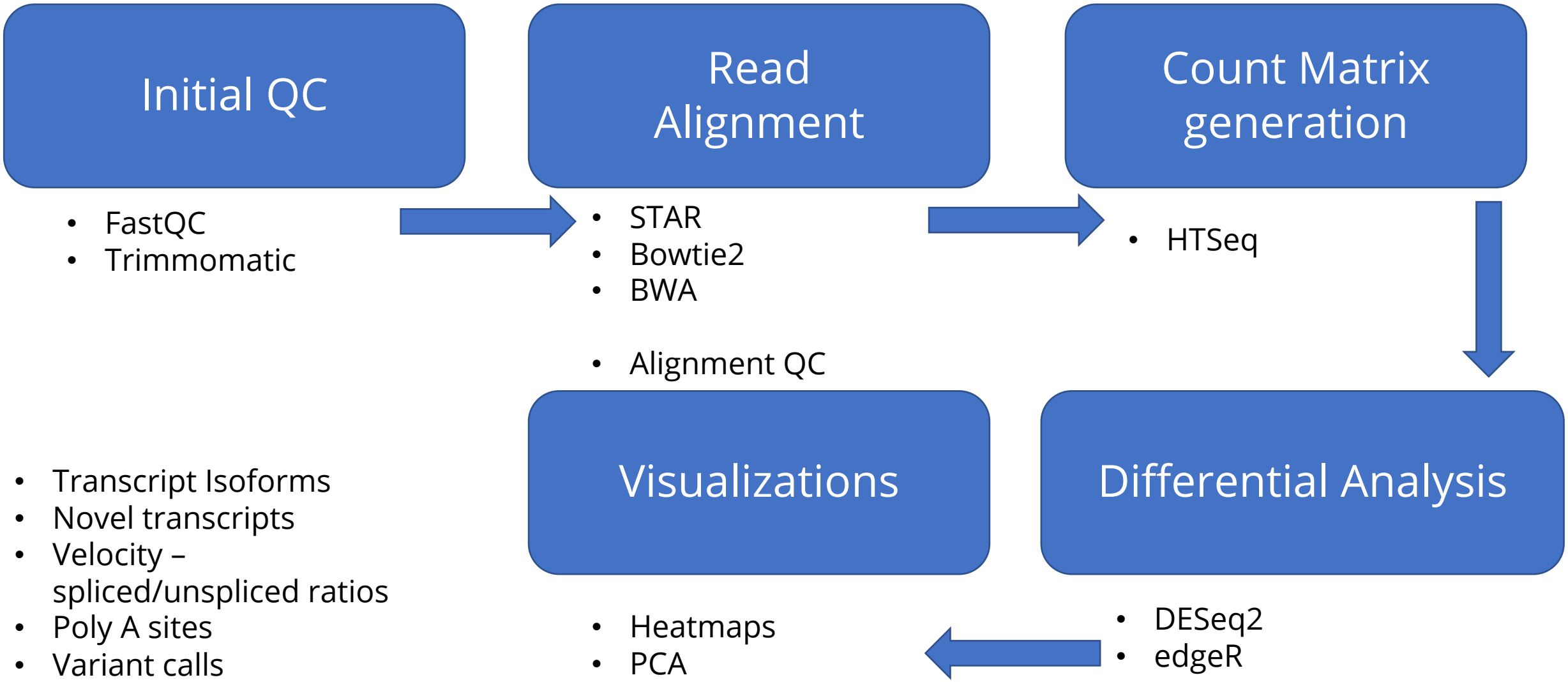
RNA-seq



RNA-seq analysis steps



RNA-seq analysis steps



Initial QC: Fastq files

- Fastq File format

Read N

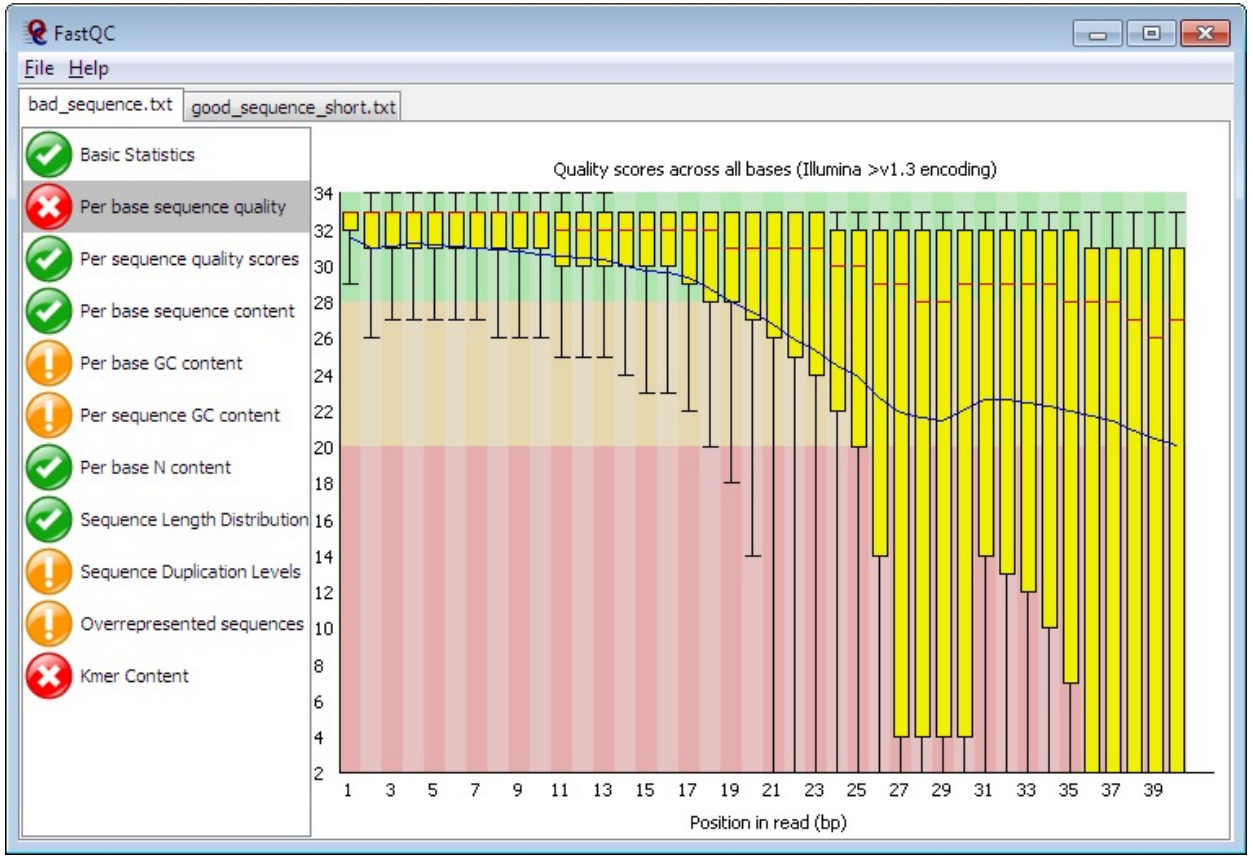
```
@K00217:70:HHJF5BBXX:1:1101:1955:1121 2:N:0:NTGCGCGC
NAAAAGAAAAGAACCCGCCGAGCAGTCAAATTCCAGAGGGCCATTACTGG
+
#A<---FJJJFJJJJ-AAJFJFA-AFJ<-FF<F7J<7--7A-7--<7-7-
```

Read N + 1

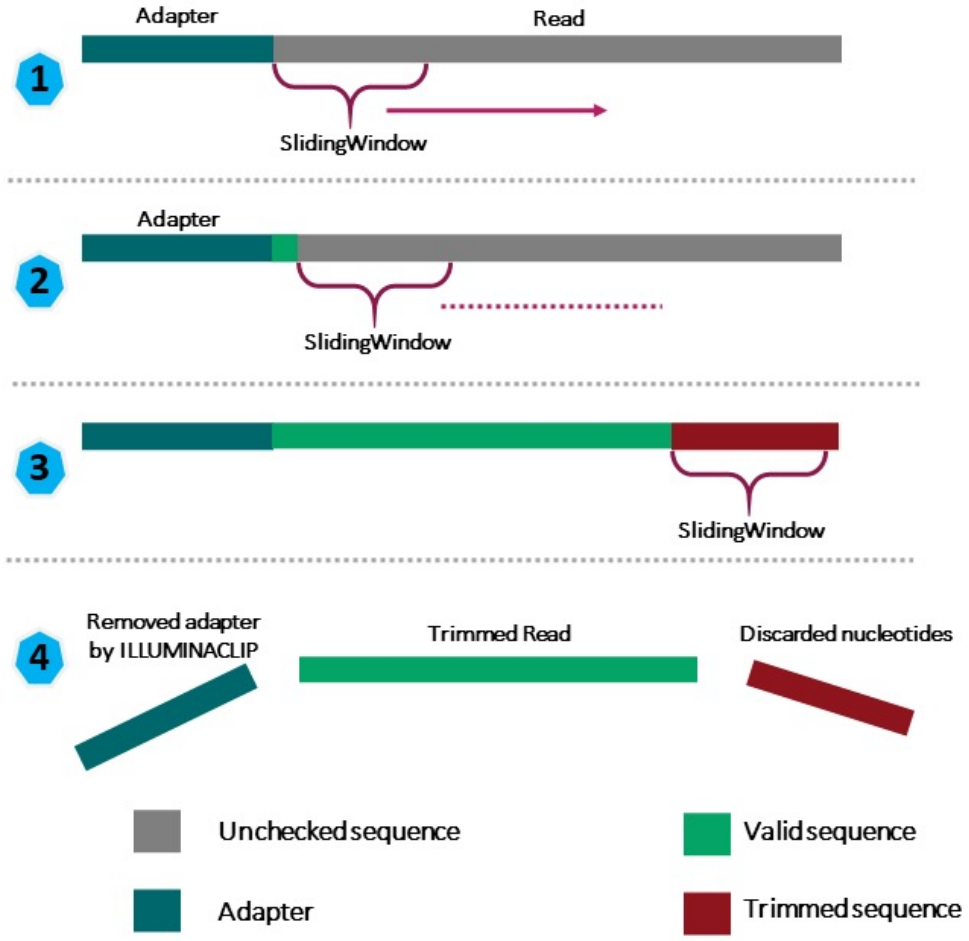
```
@K00217:70:HHJF5BBXX:1:1101:2016:1121 2:N:0:NGCGAGTA
NTCTGTCACGCACATGTGTCCTGTGGGTATAGCTAGAAGGACAGGAGGCT
+
#-<<7FJAJAFJJJJ7AJA<F-FJ<J-7F-FJJ-<AJJFFJFF-7-7--7
```

FastQC

- Quality check for Fastq files
- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/



Trimmomatic: Remove sequencing adapters



Alignment: STAR

- STAR Aligner: Fast!

Aligner	Mapping speed: million read pairs/hour		Peak physical RAM, GB	
STAR	309.2	549.9	27.0	28.4
STAR sparse	227.6	423.1	15.6	16.0
TopHat2	8.0	10.1	4.1	11.3
RUM	5.1	7.6	26.9	53.8
MapSplice	3.0	3.1	3.3	3.3
GSNAP	1.8	2.8	25.9	27.0

- Map to both genome & transcriptome
- Output: Bam files with information about genomic location of reads

Dobin et. al, 2013

Alignment QC

- STAR log

```
                                 Started job on |   Apr 23 23:17:02
                             Started mapping on |   Apr 23 23:17:04
                                    Finished on |   Apr 23 23:26:52
       Mapping speed, Million of reads per hour |   115.68

                           Number of input reads |   18894432
                       Average input read length |   298
                                     UNIQUE READS:
                    Uniquely mapped reads number |   17704240
                         Uniquely mapped reads % |   93.70%
                       Average mapped length     |   297.39
                    Number of splices: Total     |   3119841
         Number of splices: Annotated (sjdb)     |   2663436
                    Number of splices: GT/AG     |   3080422
                    Number of splices: GC/AG     |   14219
                    Number of splices: AT/AC     |   248
            Number of splices: Non-canonical     |   24952
                   Mismatch rate per base, %     |   0.49%
                      Deletion rate per base     |   0.02%
                     Deletion average length     |   2.70
                     Insertion rate per base     |   0.02%
                    Insertion average length     |   2.30
                          MULTI-MAPPING READS:
       Number of reads mapped to multiple loci   |   806405
            % of reads mapped to multiple loci   |   4.27%
       Number of reads mapped to too many loci   |   1146
            % of reads mapped to too many loci   |   0.01%
```

Count matrix generation

- Count matrix: Gene expression counts for each sample
- We need two components
  - BAM files
  - ???

Count matrix generation

- Count matrix: Gene expression counts for each sample
- We need two components
  - BAM files
  - Transcript definition!

Transcript definition: GTF files

```
1	ensembl_havana	gene	685679	686673	.	-	.	gene_id "ENSG00000284662"; gene_version "1"; gene_name "OR4F16"; gene_source "ensembl_havana"; gene_biotype "protein_coding";
1	ensembl_havana	transcript	685679	686673	.	-	.	gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; gene_name "OR4F16"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "OR4F16-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS41221"; tag "basic"; transcript_support_level "NA (assigned to previous version 3)";
1	ensembl_havana	exon	685679	686673	.	-	.	gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; exon_number "1"; gene_name "OR4F16"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "OR4F16-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS41221"; exon_id "ENSE00002324228"; exon_version "3"; tag "basic"; transcript_support_level "NA (assigned to previous version 3)";
1	ensembl_havana	CDS	685719	686654	.	-	0	gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; exon_number "1"; gene_name "OR4F16"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "OR4F16-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS41221"; protein_id "ENSP00000329982"; protein_version "2"; tag "basic"; transcript_support_level "NA (assigned to previous version 3)";
1	ensembl_havana	start_codon	686652	686654	.	-	0	gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; exon_number "1"; gene_name "OR4F16"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "OR4F16-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS41221"; tag "basic"; transcript_support_level "NA (assigned to previous version 3)";
1	ensembl_havana	stop_codon	685716	685718	.	-	0	gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; exon_number "1"; gene_name "OR4F16"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "OR4F16-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS41221"; tag "basic"; transcript_support_level "NA (assigned to previous version 3)";
1	ensembl_havana	five_prime_utr	686655	686673	.	-	.	gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; gene_name "OR4F16"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "OR4F16-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS41221"; tag "basic"; transcript_support_level "NA (assigned to previous version 3)";
1	ensembl_havana	three_prime_utr	685679	685715	.	-	.	gene_id "ENSG00000284662"; gene_version "1"; transcript_id "ENST00000332831"; transcript_version "4"; gene_name "OR4F16"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "OR4F16-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS41221"; tag "basic"; transcript_support_level "NA (assigned to previous version 3)";
1	ensembl_havana	gene	1211340	1214153	.	-	.	gene_id "ENSG00000186827"; gene_version "11"; gene_name "TNFRSF4"; gene_source "ensembl_havana"; gene_biotype "protein_coding";
1	ensembl_havana	transcript	1211340	1214153	.	-	.	gene_id "ENSG00000186827"; gene_version "11"; transcript_id "ENST00000379236"; transcript_version "4"; gene_name "TNFRSF4"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "TNFRSF4-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS11"; tag "basic"; transcript_support_level "1 (assigned to previous version 3)";
1	ensembl_havana	exon	1213983	1214153	.	-	.	gene_id "ENSG00000186827"; gene_version "11"; transcript_id "ENST00000379236"; transcript_version "4"; exon_number "1"; gene_name "TNFRSF4"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "TNFRSF4-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS11"; exon_id "ENSE00001832731"; exon_version "2"; tag "basic"; transcript_support_level "1 (assigned to previous version 3)";
1	ensembl_havana	CDS	1213983	1214127	.	-	0	gene_id "ENSG00000186827"; gene_version "11"; transcript_id "ENST00000379236"; transcript_version "4"; exon_number "1"; gene_name "TNFRSF4"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "TNFRSF4-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS11"; protein_id "ENSP00000368538"; protein_version "3"; tag "basic"; transcript_support_level "1 (assigned to previous version 3)";
1	ensembl_havana	start_codon	1214125	1214127	.	-	0	gene_id "ENSG00000186827"; gene_version "11"; transcript_id "ENST00000379236"; transcript_version "4"; exon_number "1"; gene_name "TNFRSF4"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "TNFRSF4-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS11"; tag "basic"; transcript_support_level "1 (assigned to previous version 3)";
1	ensembl_havana	exon	1213663	1213785	.	-	.	gene_id "ENSG00000186827"; gene_version "11"; transcript_id "ENST00000379236"; transcript_version "4"; exon_number "2"; gene_name "TNFRSF4"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "TNFRSF4-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS11"; exon_id "ENSE00001333068"; exon_version "1"; tag "basic"; transcript_support_level "1 (assigned to previous version 3)";
1	ensembl_havana	CDS	1213663	1213785	.	-	2	gene_id "ENSG00000186827"; gene_version "11"; transcript_id "ENST00000379236"; transcript_version "4"; exon_number "2"; gene_name "TNFRSF4"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; transcript_name "TNFRSF4-201"; transcript_source "ensembl_havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS11"; protein_id "ENSP00000368538"; protein_version "3"; tag "basic"; transcript_support_level "1 (assigned to previous version 3)";
```